BMJ Open

Design Choices for Observational Studies of the Effect of Exposure on Disease Incidence

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-031031.R1
Article Type:	Communication
Date Submitted by the Author:	08-May-2019
Complete List of Authors:	Gail, Mitchell; NCI, Biostatistics Branch Altman, Douglas G; Centre for Statistics in Medicine, Nuffield Department of Orthopaedics Cadarette, Suzanne; University of Toronto Collins, Gary; University of Oxford, Centre for Statistics in Medicine Evans, Stephen; LSHTM, Medical Statistics Unit Sekula, Peggy; Medical Center - University of Freiburg, Institute for Medical Biometry and Statistics Williamson, Elizabeth; London School of Hygiene and Tropical Medicine, Woodward, Mark; The George Institute for Global Health,
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Evidence based practice, Health services research, Occupational and environmental medicine, Public health
Keywords:	EPIDEMIOLOGY, Epidemiology < ONCOLOGY, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Cardiac Epidemiology < CARDIOLOGY

Design Choices for Observational Studies of the Effect of Exposure on Disease Incidence

Mitchell H Gail, Douglas G Altman, Suzanne M Cadarette, Gary Collins, Stephen JW Evans, Peggy Sekula, Elizabeth Williamson, Mark Woodward for the STRATOS initiative (STRengthening Analytical Thinking for Observational Studies)

May 8, 2019

4432 words in text
150 words in abstract

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA, Mitchell H. Gail senior investigator

Centre for Statistics in Medicine, University of Oxford, Oxford, UK, Douglas G. Altman professor (deceased)

Leslie Dan Faculty of Pharmacy and Dalla Lana School of Public Health, University of Toronto, Toronto, Canada, Suzanne M Cadarette associate professor

Centre for Statistics in Medicine, University of Oxford, Oxford, UK, Gary Collins professor

Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK, Stephen JW Evans professor

Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany Peggy Sekula senior scientist

Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK, Elisabeth J Williamson

The George Institute for Global Health, Oxford University, Oxford, UK and University of New South Wales, Sydney, Australia, Mark Woodward professor

Correspondence to: MH Gail gailm@mail.nih.gov Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, Room 7E138, Rockville MD 20850-9780, USA

ABSTRACT

The purpose of this paper is to help readers choose an appropriate observational study design for measuring an association between an exposure and disease incidence. We discuss cohort studies, sub-samples from cohorts (case-cohort and nested case-control designs), and population-based or hospital-based case-control studies. Good study design is the foundation of a convincing observational study. Mistakes in design are often irremediable. Key steps are understanding the scientific aims of the study and what is required to achieve them. Some designs will not yield the information required to realize the aims. The choice of design also depends on the availability of source populations and resources and requires balancing the pros and cons of various designs in view of study aims and practical constraints. We compare various cohort and case-control designs to estimate the effect of an exposure on disease incidence and mention how certain design features can reduce threats to study validity.

INTRODUCTION

Choosing an appropriate observational design to establish an association between an exposure or treatment and disease incidence is crucial to the success of the study. This paper describes design options and how to choose among them. Key points are summarized in Figure 1.

Observational studies to estimate an association between an exposure and disease incidence

In an observational study, the investigator does not control the exposure (or explanatory) variable of interest. Observational studies may be descriptive, such as studies to estimate secular trends in cancer incidence, but most assess possible causal associations. Here we focus on observational studies that *estimate an association between an exposure and disease incidence* in a particular population (the source population from which the study population was selected) over a specified time period (the risk period). Specifically, we consider cohort studies that include the entire source population or a sample from it and case-control studies that include the cases of disease and a sample of controls chosen from the same source population and risk period.

Establishing an association of an exposure with disease incidence is often a first step on the quest to establish a causal effect. Experimental studies, in which the exposure is controlled by the investigator (and may be allocated by randomization), provide strong evidence for a causal association, but are not ethical for exposures like tobacco smoking, and also may be infeasible for practical reasons. In the absence of randomization, exposures may be associated with other measured or unmeasured factors called confounders that can distort (or even hide) a true association between the exposure and health outcome or induce an apparent association when none exists. Therefore, no observational study can establish a causal relationship, but indicia, such as the strength of the association, dose response, and careful control for known confounding factors are helpful (1, 2). Usually other lines of evidence, such as laboratory experiments to establish mechanisms, are required to buttress evidence of a causal relationship.

Because observational studies often provide the only information that can be gathered ethically, it is important to design them to be as convincing and informative as possible. A chief design objective is to achieve *internal validity* by having an adequate sample size, avoiding selection biases in recruiting the study sample, measuring the exposures and outcomes accurately, controlling for confounding, and performing appropriate analyses. In addition, one often desires that the results be *generalizable* to a target population (*external validity*). Although we mention some design choices pertinent to internal and external validity, readers are encouraged to consult excellent books and papers for details (e.g.(3-12)).

Our focus is on how to choose an appropriate observational study design from among several options, namely cohort studies; subsamples from cohorts, such as case-cohort and nested case-control designs; and population-based or hospital-based case-control studies. We discuss these designs later, but we introduce them here briefly (Figure 2). In a cohort design, the cohort (study population) is obtained from the source population, baseline exposure and other covariates are measured, and cohort members are followed to determine disease incidence (Figure 2a). In the case-cohort design(13), baseline exposure and covariate information are collected from all cases and from a random sample of the entire cohort (Figure 2a). In the nested case-control design(14), baseline exposure and covariate information are collected from cases arising among the cohort members and from controls matched to each case and selected from among non-cases at risk at the time the case develops (Figure 2a). In a population-based case-control study, exposure and covariate information are collected from representative incident cases and from representative non-cases (controls) from the source population (Figure 2b). In a hospital-based case-control study, exposures from incident cases of the disease of interest (disease A in Figure 2c) are compared to exposures from incident cases of another (control) disease (B) from the same hospital (Figure 2c).

Estimating absolute risk, relative risk, absolute risk difference and relative odds of disease

To discuss these designs, we need to define measures of disease incidence and of exposure association with disease incidence for a cohort study. We are following the terminology in BMJ Best Practice at https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/how-to-calculate-risk/.

We define *absolute risk*, *relative risk*, *absolute risk difference* and *relative odds* (or *odds ratio*) by an example. Table 1 describes hypothetical outcomes for a cohort consisting of 10,000 exposed and 20,000 unexposed individuals. After 10 years of follow-up, 100 cases of disease developed among exposed and 50 among unexposed individuals. The exposure-specific absolute risks of disease were therefore 100/10,000=0.01 and 50/20,000=0.0025, respectively. The relative risk is the ratio of these absolute risks, 0.01/0.0025=4.0. The absolute risk difference is 0.01-0.0025 = 0.0075. The odds ratio (or relative odds) is the ratio of the odds of disease in exposed individuals, (100/9,900), to the odds of disease in non-exposed individuals, (50/19,950). Here the odds ratio is (100/9,900)/(50/19,950) = 4.0303.

As illustrated in Table 1, absolute risk is the probability of the disease of interest. "Risk" is sometimes used synonymously with absolute risk. Absolute risk is reduced by competing risks that kill an individual before the disease of interest develops(15). Some authors use the term absolute risk (or "pure" risk) for the risk of disease in the absence of competing mortality(15).

Suppose that an investigator retrospectively measures the exposure status of the 150 individuals with disease (cases) in Table 1 and of a random sample of 150 non-cases (or controls) from the 29,850 non-cases. The relative odds (or odds ratio) of exposure in the case-control data is expected to be (100/50)/(9900/19950)=4.0303, which equals the relative odds of disease in the cohort and is a good approximation to the relative risk, 4.0 for a rare disease(16). From these data on exposure alone, the case-control study cannot determine absolute risks, but if the disease risk in the source population is known (150/30,000=0.005 in Table 1), one can also estimate exposure-specific absolute risks (and risk differences) from case-control data(16-18).

These ideas extend to studies of time to disease onset. The hazard rate (or incidence rate) is the instantaneous rate of disease at time *t* among survivors to *t*, and the relative hazard (or hazard ratio) is the ratio of two hazard rates. The incidence rate is estimated by dividing the number of events that occur in a time interval by the corresponding cumulative time at risk of cohort members (usually expressed in person-years). From cohort data, one can estimate incidence rates as well as relative hazards(19). If one subsamples the cohort at baseline as in the case-cohort design(13), or uses a time-matched nested case-control study(14), one can estimate not only relative hazards but also exposure-specific incidence rates, exposure-specific absolute risks over a specific time interval (20), and relative risks. For further information on estimation of relative hazards from nested case-control designs, see (21-23).

A triumph of 20th century epidemiology was the demonstration of an increased risk of lung cancer in smokers. Among the most influential studies was a case-control comparison of smoking histories in lung cancer patients with those in hospitalized patients with other diseases (controls) (17). The strong relative odds found in that study was confirmed by the strong relative risks found in a later cohort study of British physicians(24, 25).

Study aims, design choices and practicalities

The appropriateness of a study design depends on the research question. If the aim is to estimate exposure-specific absolute risk, then a case-control study alone, without information on overall risk in the source population, will not provide the needed information.

Planned cohort studies are usually thought to be better than case-control studies because exposures and confounders can be reliably measured and recorded at baseline and are not subject to recall bias. However, cohort studies based on data collected routinely for other purposes, such as healthcare

utilization records, can suffer from measurement error and other threats to internal validity. Indeed, each of the designs in Tables 2 and 3 has strengths and weaknesses (Sections 3 and 4). Whether a particular design yields valid results depends on feasibility and details of study design and execution(26). Thus choosing the best design among those that can address study aims involves a context-specific balance among competing considerations(9).

DEFINING THE RESEARCH QUESTION

The most crucial aspect of study design is understanding and defining the primary research question and aims, and what is needed to achieve them. Some key issues are outlined here.

1. *How will one measure the effect of the exposure on the health outcome?* Ideally one can obtain exposure-specific absolute risks, such as 0.01 for the exposed and 0.0025 for the unexposed in Table 1. Exposure-specific absolute risks are needed to weigh the benefits and harms of an intervention, such as a program to reduce exposure or a new treatment, and some journals insist on including absolute risks whenever feasible. Often, exposure-specific incidence rates (per person-year) that take follow-up time into account are required. The relative risk and relative hazard are estimable from cohort data and approximately from case-control data via the relative odds (Section 1). Because a case-control study that collects new data can usually be conducted more quickly and cheaply than a new cohort study, estimates of relative odds and relative risks are widely used to identify risk factors for disease.

2. *What is the nature of the exposure, and how will it be measured?* The operational definition of the exposure needs to be clearly defined. If the exposure is the amount of exercise per week, this needs to be defined by protocols for a fitness-tracking device or items in a questionnaire, and if the exposure is a blood analyte, laboratory protocols for obtaining and measuring the analyte are needed. Procedures for quality control should be built into the design. To minimize artifacts from batch effects in laboratory measurements, cases and controls should be balanced within batches. If exposures are measured repeatedly in the same individuals over time, the measurement process and timing should be independent of disease status, if possible.

3. *Which confounders need to be controlled for, and how?* Control for confounding requires scientific understanding to identify risk factors for the outcome that are also possibly associated with exposure. Matched designs may enable better control for confounding (although it is still necessary to adjust for matching factors (7, 27)). Analytical methods, such as multivariable regression or propensity scoring may be used to control for confounding, provided one is able to identify and measure potential confounders.

4. *What is the target population for which results of this study might be informative?* Relative risk estimates from one population may be similar to those found in other populations. Exposure-specific absolute risks are usually more heterogeneous. For example, estimates of the absolute risk of breast cancer from BRCA1 mutations from women in families with many affected relatives are higher than absolute risks in mutation carriers from the general population(28). Thus, one should bear in mind the target population when choosing the source population and study sample.

5. *Is this a hypothesis-driven study focused on a well-defined exposure and outcome, or is it an exploratory study that examines many exposures or outcomes to discover an association?* An example of hypothesis-driven research might be to measure the association of household radon exposure with lung cancer risk(29). The designs for hypothesis-driven research should focus on such issues as the sample size needed to detect a given exposure effect and can lead to compelling evidence about an association with disease. High throughput technologies that yield thousands of measurements on a single individual make exploratory ("discovery") studies attractive. For example, comparisons of breast cancer cases and controls at hundreds of thousands of genetic loci ("genome-wide association studies") have led to the discovery of about 200 breast cancer-associated single nucleotide polymorphisms. Similarly, an exploratory cohort study of occupational formaldehyde exposure searched for mortality associations with ten lymphohematopoietic malignancies(30). Exploratory studies require statistical procedures such as Bonferroni correction to reduce false positive findings from multiple comparisons and need to be confirmed in independent data(31).

6. *Is the study large enough to provide sufficiently precise estimates of the effect of the exposure?* If confidence intervals on exposure effects are too broad, the study will not be convincing. Also, the proportion of false positive "statistically significant" findings is high in studies that are too small(32). Therefore, sample size calculations(8, 33) are needed to assure that the design meets objectives.

We focus next on hypothesis-driven studies with well-defined aims, such as: "The purpose of this study is to determine whether exposure *X* is associated with increased relative risk of disease *D*, compared to non-exposure to *X*, adjusted for confounders."

COHORTS AND SUBSAMPLES OF COHORTS

Cohort designs

The prospective cohort design provides the most general type of information on disease incidence and is easy to understand (Figure 2a, Tables 1 and 2). Cohort members without the disease of interest are

identified, exposures and covariates are recorded at date of entry into the cohort, and subsequent disease incidence is ascertained over the follow-up risk period. Related designs sub-sample a cohort (Figure 2a and Table 2). We consider dichotomous disease outcome (yes or no) over a defined time period, as in Table 1, but these ideas extend to studies of time to disease incidence. The time scale may be time since accrual into the cohort or age. The cohort study can estimate exposure-specific absolute risk (Section 2), as well as relative risks of disease and any other function of the exposure-specific absolute risk.

The prospective cohort design has several advantages in addition to its ability to estimate exposure-specific absolute risks (Table 2). First, covariates such as exposure X and potential confounders are measured at baseline, before they are influenced by the effects of incident disease. Avoidance of such "reverse causation bias" (for example, diet changes in response to incident disease) and the ability to obtain high quality exposure data at baseline are reasons for choosing this design for exposures like diet. Second, cohort studies can be designed to provide serial measurements on exposure (and other covariates) to study associations of exposure trends with disease incidence. Such cohort studies are often called *longitudinal studies*. Third, cohorts can provide data not only on the disease of primary interest but also on other diseases. Thus, a single study might provide estimates of the association of X with several diseases. Fourth, although models such as the Cox proportional hazards model(19) are often used to analyze time-to-event cohort data, many modeling approaches, such as Aalen's additive hazard model(34), can be estimated with cohort data.

The chief disadvantage of the cohort design concerns sample size and study duration for a moderately rare outcome, such as cancer incidence or stroke incidence (Table 2). The cohort needs to be large and the follow-up long to observe the number of incident cases required for sufficiently precise estimation of absolute risk or relative risk. If the exposure is also rare, such as a drug exposure or genetic mutation, even larger sample sizes are needed. The large required sample size limits the ability to capture detailed covariate information. For example, among 306,473 men and women, aged 40-73 years and followed for a median of 7.1 years in the UK Biobank Study, 287 suffered intracerebral hemorrhagic strokes(35), which is adequate to detect some associations, but not modest associations or associations with rare exposures.

It took 10 years to accumulate the cases in Table 1. One way to shorten such a study is to look for an "historical cohort" that was previously established (Table 2). For example, a mining company may have records to identify previous employees. If it were possible to retrieve information on the employees' exposures and on their previously incident health outcomes, one could analyze the cohort data without waiting for incident cases to arise. The historical cohort design may provide imperfect information,

however. Data on exposure and disease ascertainment may be incomplete. Records of who was employed may be incomplete. Unrecorded employees who stay well may remain unidentified, whereas unrecorded employees who develop disease may make health claims and be recorded as having events, which can bias incidence rates upward. Electronic health records in national databases or health maintenance organizations yield historical cohort data with information on exposures like medication use and on health outcomes but may provide limited data on confounders.

Nested case-control design

Sometimes an exposure such as a blood analyte may be too costly to measure on all members of a cohort. Blood samples may have been obtained and stored on all cohort members, but it may be much less expensive to perform the assay only on individuals who develop disease and appropriately selected controls (Table 2). For each case, the nested case-control design [14] selects r controls without replacement from among all cohort members who remain free of the disease at the time of incidence of the case. Exposure information is needed on (r+1) times the number of incident cases. Thus, in Table 1, with $N = 30,000$ people, 150 incident cases, and $r = 2$ controls per case, exposure data would be needed on 3 x 150 = 450 individuals. The nested case-control design gives valid estimates of relative hazards for studies of time to disease onset (14, 23). It rarely pays to choose more than $r = 4$ controls for each case, because the limiting factor for precise estimation of the relative hazard becomes the number of cases, not controls(36). For precise estimation of very large or small relative hazards, however, more controls are useful(37). Not only does the nested case-control design yield valid estimates of the relative hazard, but the exposure-specific absolute risk of disease may be estimated by re-weighting the control sample to the cohort population(20, 38, 39).

Case-cohort design

A potential disadvantage of the nested case-control design is that controls are time-matched to cases of a particular disease. If one wishes to study exposure associations with another type of disease, new controls will need to be chosen. The case-cohort design(13, 40) avoids this difficulty by selecting a random sub-cohort from the cohort and comparing the baseline exposures of incident cases that arise in the cohort with baseline exposures in the sub-cohort (Table 2). For example, a sub-cohort of 500 (1.67% random sample of original cohort of 30,000) might be used for comparisons against the 150 incident cases that arose in Table 1, (of whom about 1.67% x 150 = 3 are sub-cohort members). As for the nested case-control design, the success of this strategy depends on having stored blood samples (or other materials or data needed for exposure assessment) on all cohort members, but only performing the exposure

assessment on incident cases and sub-cohort members. In the previous example, exposure assessments would be required on approximately 150 + (500 – 1.67% x 150) =647 individuals, instead of 30,000. A great advantage of the case-cohort design is that the same sub-cohort can be used to study associations with several different diseases. This design also yields simple estimates of exposure-specific absolute risk as well as relative risks (Table 2).

CASE-CONTROL DESIGNS NOT NESTED IN A COHORT

Population-based case-control design

Although the nested case-control design is efficient for sampling from a well-defined cohort, often it is not possible to enumerate a suitable cohort. Nonetheless, it may be possible to obtain a random sample, or even an exhaustive sample, of all the cases that arise in a given region in a fixed time period as well as a random sample of non-cases from this source population (Figure 2b and Table 3). To avoid bias, it is important that the cases be representative of all cases and the controls be representative of all non-cases(16, 21). These population-based cases and controls constitute the study population.

The population-based case-control design is usually less expensive and time consuming than a new cohort study with primary data collection. The incident cases can be ascertained in a comparatively short time because they derive from a large source population. It is rarely necessary to sample more than $r = 4$ controls per case(36, 41).

The population-based case-control design has additional advantages. Because one can focus on a smaller number of individuals, one can obtain detailed information on possible exposures and confounders. Also, if one knows the disease incidence rate in the source population, one can estimate not only relative risks (cumulative odds ratios, incident rate ratios/relative hazards, or relative risks, depending on how the controls were sampled and rarity of disease(21)), but also exposure-specific absolute risk(16).

The population-based case-control design also has weaknesses (Table 3). First, not all the randomly selected cases and controls will agree to participate in the study, particularly if biologic specimens are required. Thus, the participating cases and controls may not be representative, and if, for example, exposed cases tend to participate more than exposed non-cases, biased odds ratios will result. Second, participants' recall of information on previous exposure and other covariates may be faulty. A particularly harmful form of misinformation on exposure is "differential recall bias," whereby cases have a different perception of previous exposures than non-cases, resulting in biased odds ratios. Studies of dietary exposures are subject to such bias, for example. Even if the exposure is based on a laboratory

measurement, a form of differential measurement error ("reverse causation") may result because the preclinical disease process may affect an individual's biochemistry or appetite, even though the biochemical feature did not cause the disease. In such circumstances, it is best to use a cohort design or a nested case-control design or case-cohort design with previously stored biologic specimens or questionnaire data. Studies of medical treatments and drug exposures are especially subject to bias from reverse causation (sometimes called "confounding by indication"), because the disease or its precursors may dictate the treatment, rather than the treatment cause the disease. This can be problematic even in cohort studies. Not all exposures are subject to biased retrospective assessment, however. For example, genotypes measured in case-control studies are not subject to recall bias or reverse causation.

Hospital-based case-control design

It may not be feasible to obtain representative population-based random samples of cases and controls if randomly selected individuals refuse to provide blood samples, for example. An alternative is to recruit cases at a hospital and to select as controls patients at the same hospital with diseases thought to be unrelated to the exposure (Figure 2c and Table 3). Cases and controls recruited in the hospital setting are likely to consent to have blood drawn for study. If the cases (disease A in Figure 2c) are representative of cases in the source population with respect to exposure and if control cases (disease B in Figure 2c) are also representative of the source population of non-cases with respect to exposure, then exposure odds ratios comparing cases to controls will be similar to those from a population-based study. However, two features of hospital-based case-control designs render them especially susceptible to bias, in addition to imperfect recall that affects all case-control designs. First, disease A cases that come to a given hospital and disease B patients that come to that hospital (and serve as controls) may not be representative of disease A cases or disease B cases in the source population, because factors such as socioeconomic status may influence who goes to a particular hospital (dotted lines in Figure 2c). Using disease B controls from the same hospital will not cause such selection biases if the selection forces act equally on patients with diseases A and B. However, this is not always true and is hard to verify. For example, the hospital may specialize in disease A, meaning that its catchment area is wide, whereas patients with the control disease B may come from near the hospital. The two groups may differ in social status, which may induce bias. The second major assumption is that the control disease B is not associated with the exposure. If the exposure is positively associated both with disease A and with disease B, the exposure odds ratios will be biased towards unity. For example, one of the first case-control studies of the association of lung cancer with smoking used patients with cardiovascular disease and with respiratory disease among the controls(17). In view of the known association of smoking with these control diseases, as is now

understood, it is likely that the odds ratios with smoking found by Doll and Hill(17), though very large, were attenuated compared to what would have been observed with population-based controls.

DISCUSSION

We emphasized the importance of defining the study aims as the key step in study design. Choosing an appropriate design requires balancing resources and study elements to best meet the study aims. For studying associations of an exposure with disease incidence, we catalogued the major design options and their strengths and weaknesses (see also (42)).

We mentioned some features of these designs that can threaten or enhance internal validity. The reader is encouraged to consult texts such as (7-9) for details. We now review these themes. Exploratory studies have special threats to internal validity because apparent associations will arise by chance if many exposures or many disease subtypes are examined. Some threats to internal validity can be mitigated by careful design. Analysis of covariate information can help control for confounding, and matched designs may facilitate and improve such analyses. Both approaches require identifying and measuring the potential confounders beforehand. Measurement error in exposure, confounders or outcome ascertainment threatens internal validity, and the study design and planning should try to reduce such errors by perfecting questionnaires, measurement instruments, and follow-up procedures. If a laboratory assay has substantial batch-to-to batch variability, then including cases and controls in each batch can reduce potential biases. Efforts to improve participation rates by those invited for a study can reduce selection biases. Missing data pose a threat to internal validity, especially if missingness is related to exposure or outcome, which will be difficult or impossible to know. Special procedures to obtain complete data on exposure and key covariates may be helpful. The design should specify the proposed analysis and required sample size to meet study objectives. Pilot studies to test the feasibility of the design and measurements are highly desirable and usually indispensable.

Even if the study is internally valid, the generalizability of the result to a target population may be questionable if the source population for the study differs from the target population. Thus, the target population needs to be considered when planning the study.

We have mentioned many factors to be considered in designing a study to estimate an association between an exposure and disease incidence. But none is more important than careful delineation of study aims and assuring that the chosen design, as outlined in Figure 2 and Tables 2 and 3, can meet those aims.

References

1. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer - recent evidence and a discussion of some questions. J Natl Cancer Inst. 1959;22(1):173-203.
2. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965;58:295-300.
3. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. 1. Principles. Am J Epidemiol. 1992;135(9):1019-28.
4. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. 2. Types of controls. Am J Epidemiol. 1992;135(9):1029-41.
5. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. 3. Design options. Am J Epidemiol. 1992;135(9):1042-50.
6. Cox DR. The design of empirical studies: towards a unified view. Eur J Epidemiol. 2016;31(3):217-28.
7. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. IARC Sci Publ. 1980(32):5-338.
8. Breslow NE, Day NE. Statistical Methods in Cancer research, Volume II: The Design and Analysis of Cohort Studies. Lyon: International Agency for Research on Cancer; 1987.
9. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Third ed. Philadelphia: Walters Kluwer | Lippincott Williams and Wilkins; 2008.
10. Woodward M. Epidemiology Study Design and Data Analysis. Third ed. Boca Raton: CRC Press Taylor and Francis Group; 2014.
11. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. International Journal of Surgery. 2014;12(12):1500-24.
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. International Journal of Surgery. 2014;12(12):1495-9.
13. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73(1):1-11.
14. Liddell FDK, McDonald JC, Thomas DC. Methods of cohort analysis - appraisal by application to asbestos mining. J R Stat Soc a Stat. 1977;140:469-91.
15. Pfeiffer RM, Gail MH. Absolute Risk: Methods and Applications in Clinical Management and Public Health. Baton Rouge: Chapman and Hall/CRC Taylor and Francis Group; 2017.
16. Cornfield J. A method of estimating comparative rates from clinical data - applications to cancer of the lung, breast and cervix. J Natl Cancer Inst. 1951;11(6):1269-75.
17. Doll R, Hill AB. Smoking and carcinoma of the lung - preliminary report. Br Med J. 1950;2(4682):739-48.
18. Gail MH. Statistics in action. Journal of the American Statistical Association. 1996;91(433):1-13.
19. Cox DR. REGRESSION MODELS AND LIFE-TABLES. Journal of the Royal Statistical Society Series B-Statistical Methodology. 1972;34(2):187-+.
20. Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. Biometrics. 1997;53(2):767-74.
21. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. Am J Epidemiol. 1982;116(3):547-53.
22. Pearce N. What does the odds ratio estimate in a case-control study. Int J Epidemiol. 1993;22(6):1189-92.
23. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika. 1978;65(1):153-8.

24. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits - A preliminary report. BMJ-British Medical Journal. 1954;1(4877):1451-5.
25. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking - A 2nd report on the mortality of British doctors. Br Med J. 1956;2(NOV10):1071-81.
26. Pearce N. Epidemiology in a changing world: variation, causation and ubiquitous risk factors. International Journal of Epidemiology. 2011;40:503-12.
27. Pearce N. Analysis of matched case-control studies. BMJ. 2016;352.
28. Antoniou A, Pharoah PDP, Narod S, Risch HA, Eyfjord JE, Hopper JL, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. Am J Hum Genet. 2003;72(5):1117-30.
29. Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, et al. Residential radon and risk of lung cancer - A combined analysis of 7 north American case-control studies. Epidemiology. 2005;16(2):137-45.
30. Beane Freeman LE, Blair A, Lubin JH, Stewart PA, Hayes RB, Hoover RN, et al. Mortality From Lymphohematopoietic Malignancies Among Workers in Formaldehyde Industries: The National Cancer Institute Cohort. J Natl Cancer Inst. 2009;101(10):751-61.
31. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med. 2014;33(11):1946-78.
32. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, et al. Design and analysis of randomized clinical-trials requiring prolonged observation of each patient. 1. Introduction and design. Br J Cancer. 1976;34(6):585-612.
33. Gail MH, Haneuse S. Power and Sample Size for Case-Control Studies. In: Borgan O, Breslow NE, Chatterjee N, Gail MH, Scott A, Wild CJ, editors. Handbook of Statistical Methods for Case-Control Studies. Boca Raton: CRC Press/Chapman and Hall; 2018. p. in press.
34. Aalen OO. A linear-regression model for the analysis of life times. Stat Med. 1989;8(8):907-25.
35. Rutten-Jacobs LC, Larsson SC, Malik R, Rannikmäe K, Sudlow CL, Dichgans M, et al. Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306 473 UK Biobank participants. BMJ. 2018;363.
36. Ury HK. Efficiency of case-control studies with multiple controls per case - continuous or dichotomous data. Biometrics. 1975;31(3):643-9.
37. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. Journal of the American Statistical Association. 1983;78(381):1-12.
38. Rivera C, Lumley T. Using the whole cohort in the analysis of countermatched samples. Biometrics. 2016;72(2):382-91.
39. Stoer NC, Samuelsen SO. Comparison of estimators in nested case-control studies with multiple outcomes. Lifetime Data Anal. 2012;18(3):261-83.
40. Kupper LL, McMichael AJ, Spirtas R. Hybrid epidemiologic study design useful in estimating relative risk. Journal of the American Statistical Association. 1975;70(351):524-8.
41. Gail M, Williams R, Byar DP, Brown C. How many controls. J Chronic Dis. 1976;29(11):723-31.
42. Borgan O., Breslow N.E., Chatterjee N, Gail M.H., Scott A., J. WC, editors. Handbook of Statistical Methods for Case-Control Studies. Boca Raton: CRC Press/Chapman and Hall; 2018

Contributors The authors are members of the design topic group of the STRATOS (Strengthening Analytical Thinking for Observational Studies) Initiative (http://www.stratos-initiative.org/), and the paper grew out of correspondence and discussions at a meeting of STRATOS. MHG, DGA, SMC, GC, SJWE, PS, EW and MW conceived the contents of the study. MHG drafted the manuscript and incorporated revisions suggested by co-authors. MHG, DGA, SMC, GC, SJWE, PS, EW and MW critically reviewed the manuscript, offered revisions, and read and approved the submitted manuscript.

Funding This paper was not supported by direct funding. GSC was supported by the NIHR Biomedical Research Centre, Oxford; MHG was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

Competing interests All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare no competing interests.

Provenance and peer review Not commissioned; externally peer reviewed.

Table 1. Numbers of incident disease cases in a cohort study of 10,000 exposed and 20,000 unexposed individuals followed for 10 years.

	Exposed	Not Exposed	Total population
Developed disease	100	50	150
Did not develop disease	9,900	19,950	29,850
	10,000	20,000	30,000

Table 2: Cohort study designs, including subsampling from the cohort

	Data needed	Estimable quantities	Strengths	Weaknesses
Prospective cohort study	Eligibility information; baseline exposure and other covariate information; dates of follow-up and diagnosis of disease(s)	Exposure-specific absolute risks; relative risks; absolute risk differences; other	Baseline exposure and other covariate data are less subject to "reverse causation" or to recall bias. Ability to obtain updated exposure values; ability to estimate absolute risks of several health outcomes	Very large samples and long-term follow-up may be needed for rare outcomes. Not feasible to obtain extensive covariate information for all members of a large cohort
Case-cohort study; sub-cohort is a subsample of the prospective cohort	As for cohort except exposure and other covariate information only needed for cases and for the subsample	As for prospective cohort	As for cohort. Expensive laboratory tests and questionnaire processing only needed for cases and members of sub-cohort. Easy to estimate absolute risks of several health outcomes.	Because one does not know at the outset who will develop disease, blood samples and unprocessed questionnaire data need to be collected (but not analyzed) for all members of the cohort. Mild loss of precision for estimating certain parameters, compared to full cohort.
Nested case-control study within a cohort; controls matched to cases on time (i.e. age or time since recruitment) from those at risk at that time	As for cohort except exposure and other covariate information only needed for cases and for the matched controls	As for prospective cohort	As for cohort. Expensive laboratory tests and questionnaire processing only needed for cases and matched controls.	As for case-cohort. Additionally, the controls are tailored to one disease.
Historical cohort study	Eligibility information; baseline exposure and other covariate information; dates of follow-up and diagnosis of disease(s). This is obtained from historical records.	As for prospective cohort	Baseline exposure and other covariate information typically not subject to "reverse causation". Because historical data are used, one does not need to wait for disease to develop.	Records (e.g. industrial administrative files) may be incomplete, making it difficult to reconstruct who was in the cohort, to obtain accurate and complete follow-up information and to obtain accurate baseline exposure and other covariate information.

Table 3. Case-control designs that are not nested within an explicit cohort

	Data needed	Estimable quantities	Strengths	Weaknesses
Population-based incident case-control study	Eligibility information; representative samples of incident cases and controls from the source population. Retrospective information on exposure and other covariates, including possible laboratory measurements.	Relative odds of disease and relative risks of disease if controls are age-matched to cases. Only if external data on disease rates in the population are available can exposure-specific absolute risk be estimated.	Few controls needed, compared to cohort study. Time to accrue cases is short, compared to cohort study. Possible to obtain extensive information on exposure and other covariates.	Exposure and other covariates subject to recall bias and reverse causation. Low participation rates may lead to biased samples of cases or controls. Usually not possible to obtain serial exposure and other covariate measurements. Usually limited to a single health outcome. However, a single large control group may serve for several diseases in a study population(40).
Hospital-based incident case-control study	Eligibility information; data from hospital cases and hospital controls with some other disease. Retrospective information on exposure and other covariates, including possible laboratory measurements.	Relative odds or relative risks with respect to the control disease(s), not necessarily with respect to the source population.	As for population-based incident case-control study. Higher participation rates than in general population and more willingness to provide biologic samples.	As for incident case-control study. Also, the cases and controls may not be representative of the general population due to selection bias for a particular hospital. If the exposure is associated with the control disease, the exposure odds ratio will be biased.

Legends

Figure 1. Key points.

Figure 2. Designs for estimating an association between an exposure and disease incidence.

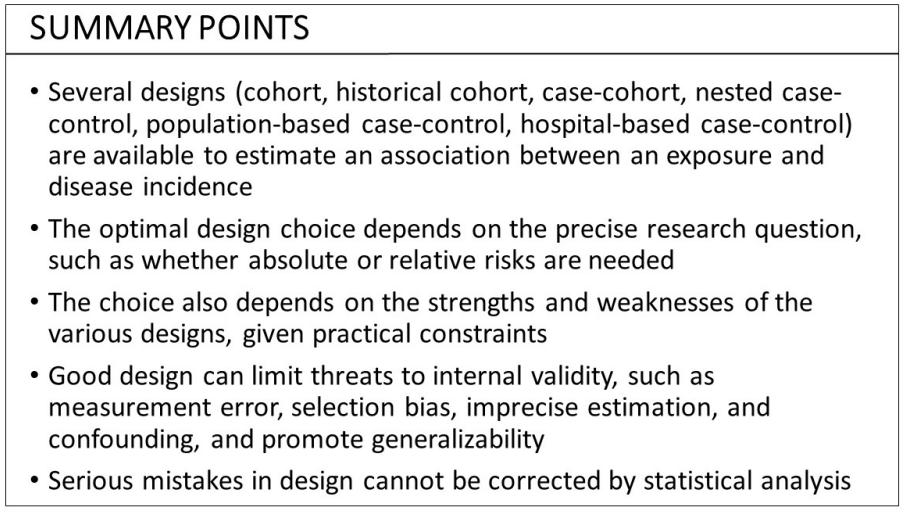

Figure 1. Key points.

338x190mm (96 x 96 DPI)

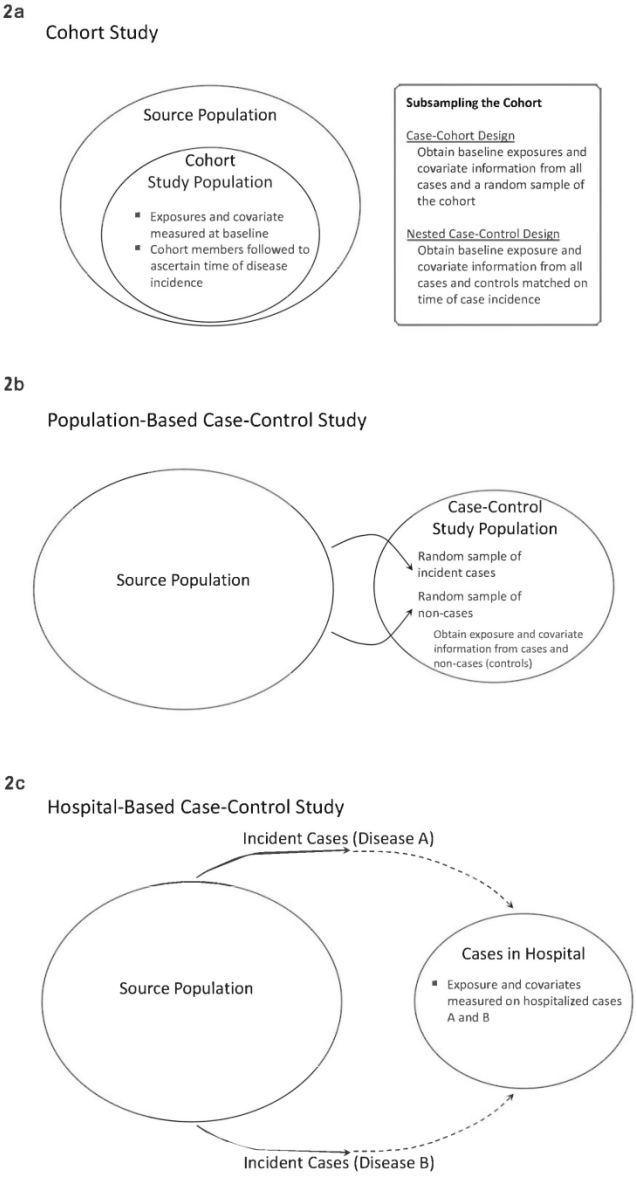

Figure 2. Designs for estimating an association between an exposure and disease incidence

215x279mm (300 x 300 DPI)

BMJ Open

Design Choices for Observational Studies of the Effect of Exposure on Disease Incidence

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-031031.R2
Article Type:	Communication
Date Submitted by the Author:	30-Aug-2019
Complete List of Authors:	Gail, Mitchell; NCI, Biostatistics Branch Altman, Douglas G; Centre for Statistics in Medicine, Nuffield Department of Orthopaedics Cadarette, Suzanne; University of Toronto Collins, Gary; University of Oxford, Centre for Statistics in Medicine Evans, Stephen; LSHTM, Medical Statistics Unit Sekula, Peggy; Medical Center - University of Freiburg, Institute for Medical Biometry and Statistics Williamson, Elizabeth; London School of Hygiene and Tropical Medicine, Woodward, Mark; The George Institute for Global Health,
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Evidence based practice, Health services research, Occupational and environmental medicine, Public health
Keywords:	EPIDEMIOLOGY, Epidemiology < ONCOLOGY, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Cardiac Epidemiology < CARDIOLOGY

Design Choices for Observational Studies of the Effect of Exposure on Disease Incidence

Mitchell H Gail, Douglas G Altman†, Suzanne M Cadarette, Gary Collins, Stephen JW Evans, Peggy Sekula, Elizabeth Williamson, and Mark Woodward for Topic Group 5 (Study Design) of the STRATOS initiative (STRengthening Analytical Thinking for Observational Studies)

August 21, 2019

5165 words in text
154 words in abstract

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA, Mitchell H. Gail senior investigator

Centre for Statistics in Medicine, University of Oxford, Oxford, UK, Douglas G. Altman professor (deceased)

Leslie Dan Faculty of Pharmacy and Dalla Lana School of Public Health, University of Toronto, Toronto, Canada, Suzanne M Cadarette associate professor

Centre for Statistics in Medicine, University of Oxford, Oxford, UK, Gary Collins professor

Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK, Stephen JW Evans professor

Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany Peggy Sekula senior scientist

Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK, Elisabeth J Williamson

The George Institute for Global Health, Oxford University, Oxford, UK and University of New South Wales, Sydney, Australia, Mark Woodward professor

Correspondence to: MH Gail gailm@mail.nih.gov Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, Room 7E138, Rockville MD 20850-9780, USA

† Deceased 3 June 1918.

ABSTRACT

The purpose of this paper is to help readers choose an appropriate observational study design for measuring an association between an exposure and disease incidence. We discuss cohort studies, sub-samples from cohorts (case-cohort and nested case-control designs), and population-based or hospital-based case-control studies. Appropriate study design is the foundation of a scientifically valid observational study. Mistakes in design are often irremediable. Key steps are understanding the scientific aims of the study and what is required to achieve them. Some designs will not yield the information required to realize the aims. The choice of design also depends on the availability of source populations and resources. Choosing an appropriate design requires balancing the pros and cons of various designs in view of study aims and practical constraints. We compare various cohort and case-control designs to estimate the effect of an exposure on disease incidence and mention how certain design features can reduce threats to study validity.

INTRODUCTION

Choosing an appropriate observational design to establish an association between an exposure or treatment and disease incidence is key to the success of the study. This paper describes design options and how to choose among them. Key points are summarized in Figure 1.

Observational studies to estimate an association between an exposure and disease incidence

In an observational study, the investigator does not control the exposure (or explanatory) variable of interest. Observational studies may be descriptive, such as studies to estimate secular trends in cancer incidence, but most assess possible causal associations. Here we focus on observational studies that *estimate an association between an exposure and disease incidence* in a particular population (the source population from which the study population was selected) over a specified time period (the risk period). Specifically, we consider cohort studies that include the entire source population or a sample from it and case-control studies that include the cases of disease and a sample of controls chosen from the same source population and risk period.

Establishing an association of an exposure with disease incidence is often a first step on the quest to establish a causal effect. Experimental studies, in which the exposure is controlled by the investigator (and may be allocated by randomization), provide strong evidence for a causal association, but are not ethical for exposures like tobacco smoking, and also may be infeasible for practical reasons. In the absence of randomization, exposures may be associated with other measured or unmeasured factors called confounders that can distort (or even hide) a true association between the exposure and health outcome or induce an apparent association when none exists. Therefore, no observational study can establish a causal relationship, but indicia, such as the strength of the association, dose response, and careful control for known confounding factors are helpful (1, 2). Usually other lines of evidence, such as laboratory experiments to establish mechanisms, are required to buttress evidence of a causal relationship.

Because observational studies often provide the only information that can be gathered ethically, it is important to design them to be as convincing and informative as possible. A chief design objective is to achieve *internal validity* by having an adequate sample size, avoiding selection biases in recruiting the study sample, measuring the exposures and outcomes accurately, controlling for confounding, and performing appropriate analyses. In addition, one often desires that the results be *generalizable* to a target population (*external validity*). Although we mention some design choices pertinent to internal and external validity, readers are encouraged to consult excellent books and papers for details (e.g.(3-12)).

The focus of this paper is on how to choose an appropriate observational study design from among several options, namely cohort studies; subsamples from cohorts, such as case-cohort and nested case-control designs; and population-based or hospital-based case-control studies. We discuss these designs later, but we introduce them here briefly (Figure 2). In a cohort design, the cohort (study population) is obtained from the source population, baseline exposure and other covariates are measured, and cohort members are followed to determine disease incidence (Figure 2a). In the case-cohort design(13), baseline exposure and covariate information are collected from all cases and from a random sample of the entire cohort (Figure 2a). In the nested case-control design(14), baseline exposure and covariate information are collected from cases arising among the cohort members and from controls time-matched to each case and selected from among non-cases at risk at the time the case develops (Figure 2a). In a population-based case-control study, exposure and covariate information are collected from representative incident cases and from representative non-cases (controls) from the source population (Figure 2b). In a hospital-based case-control study, exposures from incident cases of the disease of interest (disease A in Figure 2c) are compared to exposures from incident cases of another (control) disease (B) from the same hospital (Figure 2c).

Estimating absolute risk, relative risk, absolute risk difference and relative odds of disease

To discuss these designs, we need to define measures of disease incidence and of exposure association with disease incidence for a cohort study. *Incidence* is a measure of the probability of the occurrence of a disease in a population within a specific time period. Incidence may refer to the *incidence proportion* (also called *absolute risk*), which is the proportion of people in a population who develop disease during a specified period of time. Incidence may also refer to the *incidence rate*, which measures the occurrence of disease per unit of person-time(15). The *relative risk* is the ratio of two absolute risks, one for an exposed group and one for an unexposed group. The *absolute risk difference* is the corresponding difference in two absolute risks. The *odds of disease* corresponding to an absolute risk, AR, is AR/(1-AR). The *relative odds* (or *odds ratio*) is the ratio of the odds of disease in an exposed group to the odds of disease in an unexposed group. These definitions are consistent with the terminology in BMJ Best Practice at https://bestpractice.bmj.com/info/us/toolkit/learn-ebm/how-to-calculate-risk/.

We illustrate computation of absolute risk, relative risk, absolute risk difference and relative odds (or odds ratio) by an example. Table 1 describes hypothetical outcomes for a cohort consisting of 10,000 exposed and 20,000 unexposed individuals. After 10 years of follow-up, 100 cases of disease developed among exposed and 50 among unexposed individuals. The exposure-specific absolute risks of disease were therefore 100/10,000=0.01 and 50/20,000=0.0025, respectively. The relative risk is the ratio of

these absolute risks, 0.01/0.0025=4.0. The absolute risk difference is 0.01-0.0025 = 0.0075. The odds ratio (or relative odds) is the ratio of the odds of disease in exposed individuals, (100/9,900), to the odds of disease in non-exposed individuals, (50/19,950). Here the odds ratio is (100/9,900)/(50/19,950) = 4.0303.

As illustrated in Table 1, absolute risk is the probability of the disease of interest. "Risk" is sometimes used synonymously with absolute risk. Absolute risk is reduced by competing risks that kill an individual before the disease of interest develops(16). More generally, the competing risk can be any event that precludes subsequent observation of the event of interest. Some authors use the terms absolute risk or "pure" risk for the risk of disease in the absence of competing mortality(16).

Suppose that an investigator retrospectively measures the exposure status of the 150 individuals with disease (cases) in Table 1 and of a random sample of 150 non-cases (or controls) from the 29,850 non-cases. The relative odds (or odds ratio) of exposure in the case-control data is expected to be (100/50)/(9900/19950)=4.0303, which equals the relative odds of disease in the cohort and is a good approximation to the relative risk, 4.0 for a rare disease(17). From these data on exposure alone, the case-control study cannot determine absolute risks, but if the disease risk in the source population is known (150/30,000=0.005 in Table 1), one can also estimate exposure-specific absolute risks (and risk differences) from case-control data(17-19).

These ideas extend to studies of time to disease onset. The hazard rate (or incidence rate) is the instantaneous rate of disease at time *t* among survivors to *t*, and the relative hazard (or hazard ratio) is the ratio of two hazard rates. The incidence rate is estimated by dividing the number of events that occur in a time interval by the corresponding cumulative time at risk of cohort members (usually expressed in person-years). From cohort data, one can estimate incidence rates as well as relative hazards(20). If one subsamples the cohort at baseline as in the case-cohort design(13), or uses a time-matched nested case-control study(14), one can estimate not only relative hazards but also exposure-specific incidence rates, exposure-specific absolute risks over a specific time interval (21), and relative risks. As mentioned previously, in the time-matched nested case-control design, controls are matched to each case by sampling from among non-cases at risk at the time the case develops. For further information on estimation of relative hazards from nested case-control designs, see (22-24).

A triumph of 20th century epidemiology was the demonstration of an increased risk of lung cancer in smokers. Among the most influential studies was a case-control comparison of smoking histories in lung

cancer patients with those in hospitalized patients with other diseases (controls) (18). The strong relative odds found in that study was confirmed by the strong relative risks found in a later cohort study of British physicians(25, 26).

Study aims, design choices and practicalities

The appropriateness of a study design depends on the research question. If the aim is to estimate exposure-specific absolute risk, then a case-control study alone, without information on overall risk in the source population, will not provide the needed information.

Planned cohort studies are usually thought to be better than case-control studies because exposures and confounders can be reliably measured and recorded at baseline and are not subject to recall bias. However, cohort studies based on data collected routinely for other purposes, such as healthcare utilization records, can suffer from measurement error and other threats to internal validity. Indeed, each of the designs in Tables 2 and 3 has strengths and weaknesses (Sections 3 and 4). Whether a particular design yields valid results depends on feasibility and details of study design and execution(27).

Practical considerations include cost, time required, and access to relevant populations. Cohort studies of rare events require large samples and long follow-up. Cost or time constraints may preclude such a study. Lack of access to a relevant study population may be a factor. For example, a study of arsenic exposure in drinking water would be inefficient or futile if there was little variation of exposure in the available study population.

Thus choosing the best design among those that can address study aims involves a context-specific balance among competing considerations(9).

DEFINING THE RESEARCH QUESTION

The most crucial aspect of study design is understanding and defining the primary research question and aims, and what is needed to achieve them. Some key issues are outlined here.

1. *How will one measure the effect of the exposure on the health outcome?* Ideally one can obtain exposure-specific absolute risks, such as 0.01 for the exposed and 0.0025 for the unexposed in Table 1. Exposure-specific absolute risks are needed to weigh the benefits and harms of an intervention, such as a program to reduce exposure or a new treatment, and some journals insist on including absolute risks whenever feasible. Often, exposure-specific incidence rates (per person-year) that take follow-up time

into account are required. The relative risk and relative hazard are estimable from cohort data and approximately from case-control data via the relative odds. Because a case-control study that collects new data can usually be conducted more quickly and cheaply than a new cohort study, estimates of relative odds and relative risks are widely used to identify risk factors for disease.

2. *What is the nature of the exposure, and how will it be measured?* The operational definition of the exposure needs to be clearly defined. If the exposure is the amount of exercise per week, this needs to be defined by protocols for a fitness-tracking device or items in a questionnaire, and if the exposure is a blood analyte, laboratory protocols for obtaining and measuring the analyte are needed. Procedures for quality control should be built into the design. To minimize artifacts from batch effects in laboratory measurements, cases and controls should be balanced within batches. If exposures are measured repeatedly in the same individuals over time, the measurement process and timing should be independent of disease status, if possible.

3. *Which confounders need to be controlled for, and how?* Control for confounding requires scientific understanding to identify risk factors for the outcome that are also possibly associated with exposure. Matched designs may enable better control for confounding (although it is still necessary to adjust for matching factors (7, 28)). Analytical methods, such as multivariable regression or propensity scoring may be used to control for confounding, provided one is able to identify and measure potential confounders.

4. *What is the target population for which results of this study might be informative?* Relative risk estimates from one population may be similar to those found in other populations. Exposure-specific absolute risks are usually more heterogeneous. For example, estimates of the absolute risk of breast cancer from BRCA1 mutations from women in families with many affected relatives are higher than absolute risks in mutation carriers from the general population(29). Thus, one should bear in mind the target population when choosing the source population and study sample.

5. *Is this a hypothesis-driven study focused on a well-defined exposure and outcome, or is it an exploratory study that examines many exposures or outcomes to discover an association?* An example of hypothesis-driven research might be to measure the association of household radon exposure with lung cancer risk(30). The designs for hypothesis-driven research should focus on such issues as the sample size needed to detect a given exposure effect and can lead to compelling evidence about an association with disease. High throughput technologies that yield thousands of measurements on a single individual make exploratory ("discovery") studies attractive. For example, comparisons of breast cancer cases and

controls at hundreds of thousands of genetic loci ("genome-wide association studies") have led to the discovery of about 200 breast cancer-associated single nucleotide polymorphisms. Similarly, an exploratory cohort study of occupational formaldehyde exposure searched for mortality associations with ten lymphohematopoietic malignancies(31). Exploratory studies require statistical procedures such as Bonferroni correction to reduce false positive findings from multiple comparisons and need to be confirmed in independent data(32).

6. *Is the study large enough to provide sufficiently precise estimates of the effect of the exposure?* If confidence intervals on exposure effects are too broad, the study will not be convincing. Also, the proportion of false positive "statistically significant" findings is high in studies that are too small(33). Therefore, sample size calculations(8, 34) are needed to assure that the design meets objectives.

We focus next on hypothesis-driven studies with well-defined aims, such as: "The purpose of this study is to determine whether exposure *X* is associated with increased relative risk of disease *D*, compared to non-exposure to *X*, adjusted for confounders."

COHORTS AND SUBSAMPLES OF COHORTS

Cohort designs

The prospective cohort design provides the most general type of information on disease incidence and is easy to understand (Figure 2a, Tables 1 and 2). Cohort members without the disease of interest are identified, exposures and covariates are recorded at date of entry into the cohort, and subsequent disease incidence is ascertained over the follow-up risk period. Related designs sub-sample a cohort (Figure 2a and Table 2). We consider dichotomous disease outcome (yes or no) over a defined time period, as in Table 1, but these ideas extend to studies of time to disease incidence. The time scale may be time since accrual into the cohort or age. In studies of disease incidence, age is often used because it is strongly associated with disease incidence. In studies of death rates or disease recurrence rates following initial disease diagnosis, time since accrual (at initial diagnosis) is often used. The cohort study can estimate exposure-specific absolute risk, as well as relative risks of disease and any other function of the exposure-specific absolute risk.

The prospective cohort design has several advantages in addition to its ability to estimate exposure-specific absolute risks (Table 2). First, covariates such as exposure *X* and potential confounders are measured at baseline, before they are influenced by the effects of incident disease. Avoidance of such "reverse causation bias" (for example, diet changes in response to incident disease) and the ability to

obtain high quality exposure data at baseline are reasons for choosing this design for exposures like diet. Second, cohort studies can be designed to provide serial measurements on exposure (and other covariates) to study associations of exposure trends with disease incidence. Such cohort studies are often called *longitudinal studies*. Third, cohorts can provide data not only on the disease of primary interest but also on other diseases. Thus, a single study might provide estimates of the association of X with several diseases. Fourth, although models such as the Cox proportional hazards model(20) are often used to analyze time-to-event cohort data, many modeling approaches, such as Aalen's additive hazard model(35), can be estimated with cohort data.

The chief disadvantage of the cohort design concerns sample size and study duration for a moderately rare outcome, such as cancer incidence or stroke incidence (Table 2). The cohort needs to be large and the follow-up long to observe the number of incident cases required for sufficiently precise estimation of absolute risk or relative risk. If the exposure is also rare, such as a drug exposure or genetic mutation, even larger sample sizes are needed. The large required sample size limits the ability to capture detailed covariate information. For example, among 306,473 men and women, aged 40-73 years and followed for a median of 7.1 years in the UK Biobank Study, 287 suffered intracerebral hemorrhagic strokes(36), which is adequate to detect some associations, but not modest associations or associations with rare exposures. Because the statistical information in a cohort study of a rare event increases with the number of events observed, there can be a trade-off between study duration and the number of participants enrolled. Ten thousand participants followed for 20 years provide as much information on relative risk as 50,000 participants followed for 4 years. The longer study, however, yields data on long-term effects of exposure on absolute and relative risk. Cohort studies of events with high absolute risk, such as cancer recurrence following treatment of lung cancer, do not need to be very large or long.

Other potential limitations of cohort studies should be mentioned. It may not be feasible to collect extensive information on potential confounders in a large cohort. Because covariate information may be limited, inadequate control for confounding may yield biased estimates of relative risk. If the follow-up procedures for disease ascertainment differ between exposed and unexposed cohort members, biased estimates of relative risk may result. The available study cohort may not be representative of the general population, limiting the generalizability of the result.

It took 10 years to accumulate the cases in Table 1. One way to shorten such a study is to look for an "historical cohort" that was previously established (Table 2). For example, a mining company may have records to identify previous employees. If it were possible to retrieve information on the employees' exposures and on their previously incident health outcomes, one could analyze the cohort data without

waiting for incident cases to arise. The historical cohort design may provide imperfect information, however. Data on exposure and disease ascertainment may be incomplete. Records of who was employed may be incomplete. Unrecorded employees who stay well may remain unidentified, whereas unrecorded employees who develop disease may make health claims and be recorded as having events, which can bias incidence rates upward. Electronic health records in national databases or health maintenance organizations yield historical cohort data with information on exposures like medication use and on health outcomes but may provide limited data on confounders.

Nested case-control design

Sometimes an exposure such as a blood analyte may be too costly to measure on all members of a cohort. Blood samples may have been obtained and stored on all cohort members, but it may be much less expensive to perform the assay only on individuals who develop disease and appropriately selected controls (Figure 2a and Table 2). For each case, the nested case-control design [14] selects r controls without replacement from among all cohort members who remain free of the disease at the time of incidence of the case. Exposure information is needed on (r+1) times the number of incident cases. Thus, in Table 1, with $N = 30,000$ people, 150 incident cases, and $r = 2$ controls per case, exposure data would be needed on 3 x 150 = 450 individuals. The nested case-control design gives valid estimates of relative hazards for studies of time to disease onset (14, 24). It rarely pays to choose more than $r = 4$ controls for each case, because the limiting factor for precise estimation of the relative hazard becomes the number of cases, not controls(37). For precise estimation of very large or small relative hazards, however, more controls are useful(38). Not only does the nested case-control design yield valid estimates of the relative hazard, but the exposure-specific absolute risk of disease may be estimated by re-weighting the control sample to the cohort population(21, 39, 40).

Nested case-control studies are subject to the potential weaknesses mentioned for the full cohort except that it is feasible to analyze more baseline data to control for confounding in the nested case-control study. Nested case-control studies can also investigate associations with newly discovered analytes. These advantages can only be realized if the raw questionnaire data and biologic samples were stored for the full cohort at baseline, and if the initial informed consent or a reconsent process allowed for later investigations.

Case-cohort design

A potential disadvantage of the nested case-control design is that controls are time-matched to cases of a particular disease. If one wishes to study exposure associations with another type of disease, new controls will need to be chosen. The case-cohort design(13, 41) avoids this difficulty by selecting a random sub-cohort from the cohort and comparing the baseline exposures of incident cases that arise in the cohort with baseline exposures in the sub-cohort (Figure 2a and Table 2). For example, a sub-cohort of 500 (1.67% random sample of original cohort of 30,000) might be used for comparisons against the 150 incident cases that arose in Table 1, (of whom about 1.67% x 150 = 3 are sub-cohort members). As for the nested case-control design, the success of this strategy depends on having stored blood samples (or other materials or data needed for exposure assessment) on all cohort members, but only performing the exposure assessment on incident cases and sub-cohort members. In the previous example, exposure assessments would be required on approximately 150 + (500 – 1.67% x 150) =647 individuals, instead of 30,000. A great advantage of the case-cohort design is that the same sub-cohort can be used to study associations with several different diseases. This design also yields simple estimates of exposure-specific absolute risk as well as relative risks (Table 2).

As for the nested case-control design, baseline questionnaire data and biologic samples are needed for all cohort members, even if they will only be analyzed for incident cases and the sub-cohort, and special studies on newly discovered analytes need to be authorized by the initial informed consent or by a reconsent procedure.

CASE-CONTROL DESIGNS NOT NESTED IN A COHORT

Population-based case-control design

Although the nested case-control design is efficient for sampling from a well-defined cohort, often it is not possible to enumerate a suitable cohort. Nonetheless, it may be possible to obtain a random sample, or even an exhaustive sample, of all the incident cases that arise in a given region in a fixed time period as well as a random sample of non-cases from this source population (Figure 2b and Table 3). To avoid bias, it is important that the cases be representative of all incident cases and the controls be representative of all non-cases(17, 22). These population-based cases and controls constitute the study population.

The population-based case-control design is usually less expensive and time consuming than a new cohort study with primary data collection. The incident cases can be ascertained in a comparatively short time because they derive from a large source population. It is rarely necessary to sample more than $r = 4$ controls per case(37, 42).

The population-based case-control design has additional advantages. Because one can focus on a smaller number of individuals, one can obtain detailed information on possible exposures and confounders. Also, if one knows the disease incidence rate in the source population, one can estimate not only relative risks (cumulative odds ratios, incident rate ratios/relative hazards, or relative risks, depending on how the controls were sampled and rarity of disease(22)), but also exposure-specific absolute risk(17).

The population-based case-control design also has weaknesses (Table 3). First, absolute risk cannot be estimated unless external information on disease incidence in the source population is available. Second, not all the randomly selected cases and controls will agree to participate in the study, particularly if biologic specimens are required. Thus, the participating cases and controls may not be representative, and if, for example, exposed cases tend to participate more than exposed non-cases, biased odds ratios will result. Third, participants' recall of information on previous exposure and other covariates may be faulty. A particularly harmful form of misinformation on exposure is "differential recall bias," whereby cases have a different perception of previous exposures than non-cases, resulting in biased odds ratios. Studies of dietary exposures are subject to such bias, for example. Even if the exposure is based on a laboratory measurement, a form of differential measurement error ("reverse causation") may result because the preclinical disease process may affect an individual's biochemistry or appetite, even though the biochemical feature did not cause the disease. In such circumstances, it is best to use a cohort design or a nested case-control design or case-cohort design with previously stored biologic specimens or questionnaire data. Studies of medical treatments and drug exposures are especially subject to bias from reverse causation (sometimes called “confounding by indication”), because the disease or its precursors may dictate the treatment, rather than the treatment affect the disease. This can be problematic even in cohort studies. Not all exposures are subject to biased retrospective assessment, however. For example, genotypes measured in case-control studies are not subject to recall bias or reverse causation.

Sometimes a case-control study includes prevalent as well as incident cases. A prevalent case is a person whose disease developed before the study began and who survived to the beginning of the study. If the exposure of interest for disease incidence also affects survival following disease incidence, estimates of relative risks for incidence can be distorted by inclusion of the prevalent cases. Because the relative risk of disease incidence is a key parameter for studying disease etiology, prevalent cases should be excluded or used with caution in such studies(43).

Hospital-based case-control design

It may not be feasible to obtain representative population-based random samples of cases and controls if randomly selected individuals refuse to provide blood samples, for example. An alternative is to recruit cases at a hospital and to select as controls patients at the same hospital with diseases thought to be unrelated to the exposure (Figure 2c and Table 3). Cases and controls recruited in the hospital setting are likely to consent to have blood drawn for study. If the cases (disease A in Figure 2c) are representative of cases in the source population with respect to exposure and if control cases (disease B in Figure 2c) are also representative of the source population of non-cases with respect to exposure, then exposure odds ratios comparing cases to controls will be similar to those from a population-based study. However, two features of hospital-based case-control designs render them especially susceptible to bias, in addition to imperfect recall that affects all case-control designs. First, disease A cases that come to a given hospital and disease B patients that come to that hospital (and serve as controls) may not be representative of disease A cases or disease B cases in the source population, because factors such as socioeconomic status may influence who goes to a particular hospital (dotted lines in Figure 2c). Using disease B controls from the same hospital will not cause such selection biases if the selection forces act equally on patients with diseases A and B. However, this is not always true and is hard to verify. For example, the hospital may specialize in disease A, meaning that its catchment area is wide, whereas patients with the control disease B may come from near the hospital. The two groups may differ in social status, which may induce bias. The second major assumption is that the control disease B is not associated with the exposure. If the exposure is positively associated both with disease A and with disease B, the exposure odds ratios will be biased towards unity. For example, one of the first case-control studies of the association of lung cancer with smoking used patients with cardiovascular disease and with respiratory disease among the controls(18). In view of the known association of smoking with these control diseases, as is now understood, it is likely that the odds ratios with smoking found by Doll and Hill(18), though very large, were attenuated compared to what would have been observed with population-based controls.

Another weakness of hospital-based case-control studies is that they do not yield estimates of absolute risk (Table 3).

DISCUSSION

We emphasized the importance of defining the study aims as the key step in study design. Choosing an appropriate design requires balancing resources and study elements to best meet the study aims. For studying associations of an exposure with disease incidence, we catalogued the major design options and their strengths and weaknesses (see also (44)).

We mentioned some features of these designs that can threaten or enhance internal validity. The reader is encouraged to consult texts such as (7-9) for details. We now review these themes. Exploratory studies have special threats to internal validity because apparent associations will arise by chance if many exposures or many disease subtypes are examined. Some threats to internal validity can be mitigated by careful design. Analysis of covariate information can help control for confounding, and matched designs may facilitate and improve such analyses. Both approaches require identifying and measuring the potential confounders beforehand. Measurement error in exposure, confounders or outcome ascertainment threatens internal validity, and the study design and planning should try to reduce such errors by perfecting questionnaires, measurement instruments, and follow-up procedures. If a laboratory assay has substantial batch-to-to batch variability, then including cases and controls in each batch can reduce potential biases. Efforts to improve participation rates by those invited for a study can reduce selection biases. Missing data pose a threat to internal validity, especially if missingness is related to exposure or outcome, which will be difficult or impossible to know. Special procedures to obtain complete data on exposure and key covariates may be helpful. The design should specify the proposed analysis and required sample size to meet study objectives. Pilot studies to test the feasibility of the design and measurements are highly desirable and usually indispensable.

Even if the study is internally valid, the generalizability of the result to a target population may be questionable if the source population for the study differs from the target population. Thus, the target population needs to be considered when planning the study.

We have mentioned many factors to be considered in designing a study to estimate an association between an exposure and disease incidence. But none is more important than careful delineation of study aims and assuring that the chosen design, as outlined in Figure 2 and Tables 2 and 3, can meet those aims.

Legends

Figure 1 Key points.

Figure 2 Designs for estimating an association between an exposure and disease incidence.

Footnotes

Contributors

MHG, DGA, SMC, GC, SJWE, PS, EW and MW conceived the contents of the study. MHG drafted the manuscript. DGA, SMC, GC, SJWE, PS, EW and MW critically reviewed and edited it. MHG, SMC, GC, SJWE, PS, EW and MW gave final approval of the version to be published

and are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. DGA died during the preparation of the manuscript. MHG is the guarantor.

Funding
GSC was supported by the NIHR Biomedical Research Centre, Oxford; MHG was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data available.

Patient consent for publication Not required.

Acknowledgements
Acknowledgements: The authors are members of the Topic Group 5 (Study Design) of the STRATOS (STRengthening Analytical Thinking for Observational Studies) Initiative (http://www.stratos-initiative.org/). This Topic Group included Suzanne M Cadarette, Gary Collins, Stephen JW Evans, Mitchell H Gail, Neil Pearce, Peggy Sekula, Elizabeth Williamson, and Mark Woodward at the time this paper was developed as part of the STRATOS Initiative, whose objective is to provide accessible and accurate guidance in the design and analysis of observational studies. No patients or members of the public were involved in the creation of this article.

References

1. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer - recent evidence and a discussion of some questions. J Natl Cancer Inst. 1959;22(1):173-203.
2. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965;58:295-300.
3. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. 1. Principles. Am J Epidemiol. 1992;135(9):1019-28.
4. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. 2. Types of controls. Am J Epidemiol. 1992;135(9):1029-41.
5. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. 3. Design options. Am J Epidemiol. 1992;135(9):1042-50.
6. Cox DR. The design of empirical studies: towards a unified view. Eur J Epidemiol. 2016;31(3):217-28.
7. Breslow NE, Day NE. Statistical methods in cancer research. Volume I - The analysis of case-control studies. IARC Sci Publ. 1980(32):5-338.
8. Breslow NE, Day NE. Statistical Methods in Cancer research, Volume II: The Design and Analysis of Cohort Studies. Lyon: International Agency for Research on Cancer; 1987.
9. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Third ed. Philadelphia: Walters Kluwer | Lippincott Williams and Wilkins; 2008.
10. Woodward M. Epidemiology Study Design and Data Analysis. Third ed. Boca Raton: CRC Press Taylor and Francis Group; 2014.
11. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. International Journal of Surgery. 2014;12(12):1500-24.
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. International Journal of Surgery. 2014;12(12):1495-9.
13. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73(1):1-11.
14. Liddell FDK, McDonald JC, Thomas DC. Methods of cohort analysis - appraisal by application to asbestos mining. J R Stat Soc a Stat. 1977;140:469-91.
15. Rothman KJ, Greenland S. Modern Epidemiology. Philadelphia: Lippincott-Raven; 1998.
16. Pfeiffer RM, Gail MH. Absolute Risk: Methods and Applications in Clinical Management and Public Health. Baton Rouge: Chapman and Hall/CRC Taylor and Francis Group; 2017.
17. Cornfield J. A method of estimating comparative rates from clinical data - applications to cancer of the lung, breast and cervix. J Natl Cancer Inst. 1951;11(6):1269-75.
18. Doll R, Hill AB. Smoking and carcinoma of the lung - preliminary report. Br Med J. 1950;2(4682):739-48.
19. Gail MH. Statistics in action. Journal of the American Statistical Association. 1996;91(433):1-13.
20. Cox DR. REGRESSION MODELS AND LIFE-TABLES. Journal of the Royal Statistical Society Series B-Statistical Methodology. 1972;34(2):187-+.
21. Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. Biometrics. 1997;53(2):767-74.
22. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. Am J Epidemiol. 1982;116(3):547-53.
23. Pearce N. What does the odds ratio estimate in a case-control study. Int J Epidemiol. 1993;22(6):1189-92.
24. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika. 1978;65(1):153-8.

25. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits - A preliminary report. BMJ-British Medical Journal. 1954;1(4877):1451-5.
26. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking - A 2nd report on the mortality of British doctors. Br Med J. 1956;2(NOV10):1071-81.
27. Pearce N. Epidemiology in a changing world: variation, causation and ubiquitous risk factors. International Journal of Epidemiology. 2011;40:503-12.
28. Pearce N. Analysis of matched case-control studies. BMJ. 2016;352.
29. Antoniou A, Pharoah PDP, Narod S, Risch HA, Eyfjord JE, Hopper JL, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. Am J Hum Genet. 2003;72(5):1117-30.
30. Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, et al. Residential radon and risk of lung cancer - A combined analysis of 7 north American case-control studies. Epidemiology. 2005;16(2):137-45.
31. Beane Freeman LE, Blair A, Lubin JH, Stewart PA, Hayes RB, Hoover RN, et al. Mortality From Lymphohematopoietic Malignancies Among Workers in Formaldehyde Industries: The National Cancer Institute Cohort. J Natl Cancer Inst. 2009;101(10):751-61.
32. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med. 2014;33(11):1946-78.
33. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, et al. Design and analysis of randomized clinical-trials requiring prolonged observation of each patient. 1. Introduction and design. Br J Cancer. 1976;34(6):585-612.
34. Gail MH, Haneuse S. Power and Sample Size for Case-Control Studies. In: Borgan O, Breslow NE, Chatterjee N, Gail MH, Scott A, Wild CJ, editors. Handbook of Statistical Methods for Case-Control Studies. Boca Raton: CRC Press/Chapman and Hall; 2018. p. in press.
35. Aalen OO. A linear-regression model for the analysis of life times. Stat Med. 1989;8(8):907-25.
36. Rutten-Jacobs LC, Larsson SC, Malik R, Rannikmäe K, Sudlow CL, Dichgans M, et al. Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306 473 UK Biobank participants. BMJ. 2018;363.
37. Ury HK. Efficiency of case-control studies with multiple controls per case - continuous or dichotomous data. Biometrics. 1975;31(3):643-9.
38. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. Journal of the American Statistical Association. 1983;78(381):1-12.
39. Rivera C, Lumley T. Using the whole cohort in the analysis of countermatched samples. Biometrics. 2016;72(2):382-91.
40. Stoer NC, Samuelsen SO. Comparison of estimators in nested case-control studies with multiple outcomes. Lifetime Data Anal. 2012;18(3):261-83.
41. Kupper LL, McMichael AJ, Spirtas R. Hybrid epidemiologic study design useful in estimating relative risk. Journal of the American Statistical Association. 1975;70(351):524-8.
42. Gail M, Williams R, Byar DP, Brown C. How many controls. J Chronic Dis. 1976;29(11):723-31.
43. Begg CB, Gray RJ. Methodology for case-control studies with prevalent cases. Biometrika. 1987;74(1):191-5.
44. Borgan O., Breslow N.E., Chatterjee N, Gail M.H., Scott A., J. WC, editors. Handbook of Statistical Methods for Case-Control Studies. Boca Raton: CRC Press/Chapman and Hall; 2018

Table 1. Numbers of incident disease cases in a cohort study of 10,000 exposed and 20,000 unexposed individuals followed for 10 years.

	Exposed	Not Exposed	Total population
Developed disease	100	50	150
Did not develop disease	9,900	19,950	29,850
	10,000	20,000	30,000

Table 2: Cohort study designs, including subsampling from the cohort

	Data needed	Quantities that can be estimated	Strengths	Weaknesses
Prospective cohort study	Eligibility information; baseline exposure and other covariate information; dates of follow-up and diagnosis of disease(s)	Exposure-specific absolute risks; relative risks; absolute risk differences; other	Baseline exposure and other covariate data are less subject to "reverse causation" or to recall bias. Ability to obtain updated exposure values; ability to estimate absolute risks of several health outcomes	Very large samples and long-term follow-up may be needed for rare outcomes. Not feasible to obtain extensive covariate information for all members of a large cohort. Potential selection biases. Potential differential follow-up by exposure group.
Case-cohort study; sub-cohort is a subsample of the prospective cohort	As for cohort except exposure and other covariate information only needed for cases and for the subsample	As for prospective cohort	As for cohort. Expensive laboratory tests and questionnaire processing only needed for cases and members of sub-cohort. Easy to estimate absolute risks of several health outcomes.	Because one does not know at the outset who will develop disease, blood samples and unprocessed questionnaire data need to be collected (but not analyzed) for all members of the cohort. Mild loss of precision for estimating certain parameters, compared to full cohort.
Nested case-control study within a cohort; controls matched to cases on time (i.e. age or time since recruitment) from those at risk at that time	As for cohort except exposure and other covariate information only needed for cases and for the matched controls	As for prospective cohort	As for cohort. Expensive laboratory tests and questionnaire processing only needed for cases and matched controls.	As for case-cohort. Additionally, the controls are tailored to one disease.
Historical cohort study	Eligibility information; baseline exposure and other covariate information; dates of follow-up and diagnosis of disease(s). This is obtained from historical records.	As for prospective cohort	Baseline exposure and other covariate information typically not subject to "reverse causation". Because historical data are used, one does not need to wait for disease to develop.	Records (e.g. industrial administrative files) may be incomplete, making it difficult to reconstruct who was in the cohort, to obtain accurate and complete follow-up information and to obtain accurate baseline exposure and other covariate information.

Table 3. Case-control designs that are not nested within an explicit cohort

	Data needed	Quantities that can be estimated	Strengths	Weaknesses
Population-based incident case-control study	Eligibility information; representative samples of incident cases and controls from the source population. Retrospective information on exposure and other covariates, including possible laboratory measurements.	Relative odds of disease and relative risks of disease if controls are age-matched to cases. Only if external data on disease rates in the population are available can exposure-specific absolute risk be estimated.	Few controls needed, compared to cohort study. Time to accrue cases is short, compared to cohort study. Possible to obtain extensive information on exposure and other covariates.	Exposure and other covariates subject to recall bias and reverse causation. Low participation rates may lead to biased samples of cases or controls. Usually not possible to obtain serial exposure and other covariate measurements. Usually limited to a single health outcome. However, a single large control group may serve for several diseases in a study population(41).
Hospital-based incident case-control study	Eligibility information; data from hospital cases and hospital controls with some other disease. Retrospective information on exposure and other covariates, including possible laboratory measurements.	Relative odds or relative risks with respect to the control disease(s), not necessarily with respect to the source population.	As for population-based incident case-control study. Higher participation rates than in general population and more willingness to provide biologic samples.	As for incident case-control study. Also, the cases and controls may not be representative of the general population due to selection bias for a particular hospital. If the exposure is associated with the control disease, the exposure odds ratio will be biased.

SUMMARY POINTS

- Several designs (cohort, historical cohort, case-cohort, nested case-control, population-based case-control, hospital-based case-control) are available to estimate an association between an exposure and disease incidence
- The optimal design choice depends on the precise research question, such as whether absolute or relative risks are needed
- The choice also depends on the strengths and weaknesses of the various designs, given practical constraints
- Good design can limit threats to internal validity, such as measurement error, selection bias, imprecise estimation, and confounding, and promote generalizability
- Serious mistakes in design cannot be corrected by statistical analysis

Figure 1. Key points.

338x190mm (96 x 96 DPI)

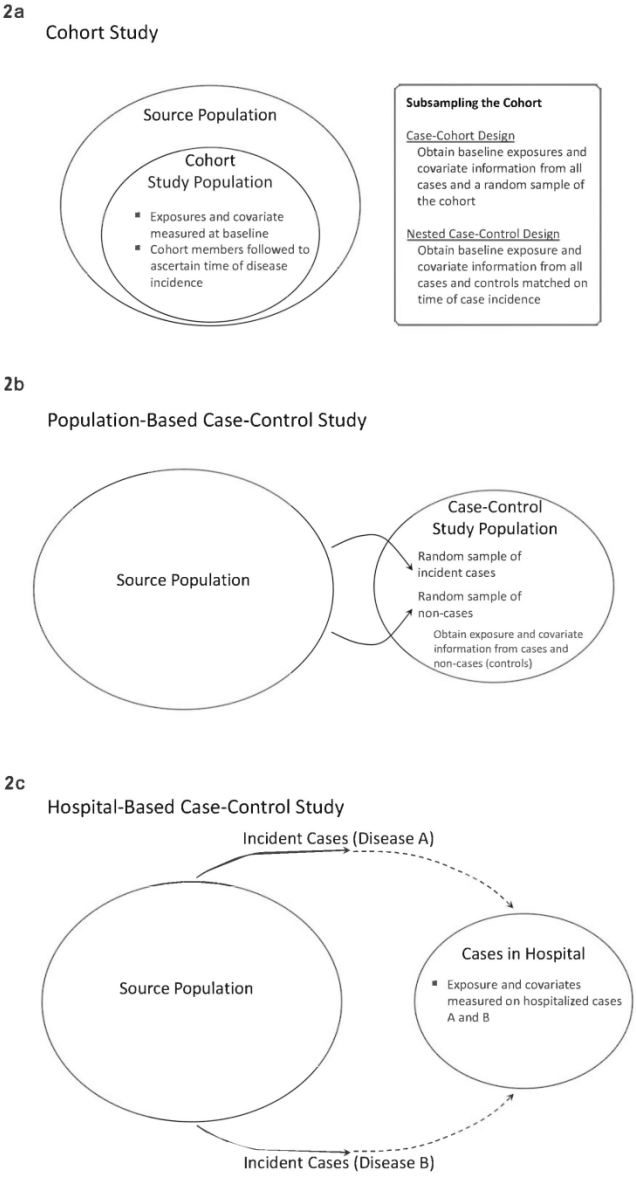

Figure 2. Designs for estimating an association between an exposure and disease incidence

215x279mm (300 x 300 DPI)